

Statistical Delay Tradeoffs in Buffer-Aided Two-Hop Wireless Communication Systems

Deli Qiao and M. Cenk Gursoy

Abstract

This paper analyzes the impact of statistical delay constraints on the achievable rate of a two-hop wireless communication link, in which the communication between a source and a destination is accomplished via a buffer-aided relay node. It is assumed that there is no direct link between the source and the destination, and the buffer-aided relay forwards the information to the destination by employing the decode-and-forward scheme. Given statistical delay constraints specified via maximum delay and delay violation probability, the tradeoff between the statistical delay constraints imposed on any two concatenated queues is identified. With this characterization, the maximum constant arrival rates that can be supported by this two-hop link are obtained by determining the effective capacity of such links as a function of the statistical delay constraints, signal-to-noise ratios (SNR) at the source and relay, and the fading distributions of the links. It is shown that asymmetric statistical delay constraints at the buffers of the source and relay node can improve the achievable rate. Overall, the impact of the statistical delay tradeoff on the achievable throughput is provided.

Index Terms

Two-hop wireless links, statistical delay constraints, quality of service (QoS) constraints, fading channels, effective capacity, delay violation probability, full-duplex relaying.

I. INTRODUCTION

With the widespread use of smart-phones and tablets, the volume of global mobile traffic has increased explosively in recent years. The portion of multimedia data, such as mobile video and voice

D. Qiao is with the School of Information Science and Technology, East China Normal University, Shanghai, China, 200241 (e-mail: dlqiao@ce.ecnu.edu.cn). M. Cenk Gursoy is with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 (email: mcgursoy@syr.edu)

This work was supported in part by the National Natural Science Foundation of China under Grants (61571191, 61572192). The material in this paper has been presented in part at the 2015 IEEE Global Communications Conference (Globecom), San Diego, United States, Dec 2015.

over IP (VoIP), has surged significantly within this wireless traffic [1]. In such multimedia traffic, delay is an important consideration. Meanwhile, providing deterministic quality of service (QoS) guarantees is challenging in wireless systems, since the instantaneous rate of the channel varies randomly depending on numerous factors, such as mobility, changing environment and multipath fading [2]. Therefore, providing statistical QoS guarantees is more suitable in such randomly-varying wireless environment.

Effective bandwidth theory has been developed to analyze high-speed communication systems operating under statistical queueing constraints [3], [4]. The queueing constraints are imposed on buffer violation probabilities and are specified by the QoS exponent θ , which dictates the exponential decay rate of the queue length in the stable state. Also, Chang and Zang have characterized the effective bandwidths of time-varying departure processes in [5], which can be utilized to analyze the volatile wireless systems. Moreover, Wu and Negi in [6] defined the dual concept of effective capacity, which provides the maximum constant arrival rate that can be supported by a given departure process while satisfying statistical delay constraints. The analysis and application of effective capacity in various settings have attracted much interest recently (see e.g. [7]-[20] and references therein).

In this paper, we study the achievable rate of two-hop systems operating under statistical delay constraints. In particular, we assume that there are buffers at both the source and the relay nodes, and consider the queueing delay introduced by the buffers. Note that [12]-[20] have also recently investigated the effective capacity of the relay channels. For instance, Tang and Zhang in [12] analyzed the power allocation policies of relay networks, where the relay node is assumed to have no queue, i.e., the packets arriving to the relay node are forwarded immediately. In [13], Liu *et al.* considered the cooperation of two users for data transmission, where the interchanged data goes through only the queue of the other user. Parag and Chamberland in [14] provided a queueing analysis of a butterfly network with constant rate for each link, while assuming that there is no congestion at the intermediate nodes. The effective capacity of the two-hop link in the presence of the statistical queueing constraints at the source and relay node is given in [15], and the performance for multi-relay links is analyzed in [16].

In this work, as a significant departure from previous works, we consider statistical end-to-end delay constraints, imposed as the limitations on the maximum delay and delay violation probability. Note that statistical end-to-end delay analysis can also be found in [17]-[20]. In [17], Wu and Negi considered statistical end-to-end delay constraints for half-duplex relays, and gave an effective

capacity formulation with time allocation to the different hops. In [18]-[20], the authors considered the statistical end-to-end delay constraints of multi-hop links, while assuming that the statistical delay violation probability of the queues are equal. However, it is possible that the relay can tolerate more stringent delay constraints while not affecting the system performance [15]. Therefore, we seek to determine the optimal statistical QoS exponents of the buffers under given *end-to-end* delay constraints. Additionally, we note that the analysis of buffer-aided systems have attracted much interest recently (see e.g., [21]-[24] and reference therein). In such analysis, the authors considered the case that only the relay node has buffer, and the average queueing delay is investigated [22]. The contributions can be summarized as follows:

- 1) We characterize the tradeoff between the statistical delay constraints at the source and relay nodes, providing a framework for dynamically adjusting the delay constraints of any two interacting queues.
- 2) With the identified interplay, we then derive the effective capacity of the two-hop links under a target statistical end-to-end delay constraint by optimizing over the statistical queueing constraints at the queues of the source and relay nodes.
- 3) We also describe a method for obtaining the effective capacity in such settings. Additionally, we show that symmetric delay constraints at the two buffers do not always lead to the optimal performance. Instead, asymmetric delay constraints, e.g., when the delay constraint at one queue is more relaxed, can lead to larger achievable rates for the two-hop system, which we verify via numerical results. Moreover, it is demonstrated that the improvement is affected by the statistical delay constraints, the signal-to-noise ratio (SNR) levels and the channel conditions of the links.

The rest of this paper is organized as follows. In Section II, the system model and necessary preliminaries are described. In Section III, we present the tradeoff between the statistical delay constraints of any two concatenated queues. We describe our main results for block-fading channels in Section IV, with numerical results provided in Section V. Finally, in Section VI, we conclude the paper.

II. PRELIMINARIES

A. System Model

The two-hop communication link is depicted in Figure 1. In this model, source **S** is sending information to the destination **D** with the help of the intermediate relay node **R**. We assume that

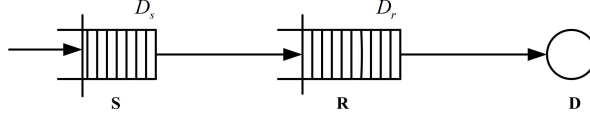


Fig. 1. The system model.

there is no direct link between S and D (which, for instance, holds, if these nodes are sufficiently far apart in distance). Both the source and the intermediate relay nodes are equipped with buffers. Hence, for the information flow of such links, the queueing delay experienced is given by $D = D_s + D_r$, where D_s and D_r denote the stationary delay experienced in the queue at the source and relay node, respectively.

We consider a full-duplex relay, and hence assume that reception and transmission can be performed simultaneously at the relay node. Note that full-duplex relaying can be achieved through some form of analog self-interference cancellation followed by digital self-interference cancellation in the baseband domain [28], [29]. In the i th symbol duration, the signal Y_r received at the relay from the source and the signal Y_d received at the destination from the relay can be expressed as

$$Y_r[i] = g_1[i]X_1[i] + n_1[i], \quad (1)$$

$$Y_d[i] = g_2[i]X_2[i] + n_2[i], \quad (2)$$

where X_j for $j = \{1, 2\}$ denote the inputs for the links S – R and R – D, respectively. More specifically, X_1 is the signal sent from the source and X_2 is sent from the relay. The inputs are subject to individual average energy constraints $\mathbb{E}\{|X_j|^2\} \leq \bar{P}_j/B, j = \{1, 2\}$ where B is the bandwidth. Assuming that the symbol rate is B complex symbols per second, we can easily see that the symbol energy constraint of \bar{P}_j/B implies that the channel input has a power constraint of \bar{P}_j . We assume that the fading coefficients $g_j, j = \{1, 2\}$ are jointly stationary and ergodic discrete-time processes, and we denote the magnitude-square of the fading coefficients by $z_j[i] = |g_j[i]|^2$. Above, in the channel input-output relationships, the noise component $n_j[i]$ is a zero-mean, circularly symmetric, complex Gaussian random variable with variance $\mathbb{E}\{|n_j[i]|^2\} = N_j$ for $j = 1, 2$. The additive Gaussian noise samples $\{n_j[i]\}$ are assumed to form an independent and identically distributed (i.i.d.) sequence. We denote the signal-to-noise ratios as $\text{SNR}_j = \frac{\bar{P}_j}{N_j B}$.

B. Statistical Delay Constraints

Suppose that the queue is stable and there exists a unique $\theta > 0$ such that

$$\Lambda_A(\theta) + \Lambda_C(-\theta) = 0, \quad (3)$$

where $\Lambda_A(\theta)$ and $\Lambda_C(\theta)$ are the logarithmic moment generating functions (LMGFs) of the arrival and service processes, respectively. Then, [5]

$$\lim_{Q_{\max} \rightarrow \infty} \frac{\log \Pr\{Q > Q_{\max}\}}{Q_{\max}} = -\theta. \quad (4)$$

where Q is the stationary queue length. Throughout the text, logarithm expressed without a base, i.e., $\log(\cdot)$, refers to the natural logarithm $\log_e(\cdot)$.

We need to guarantee that the statistical delay performance of the two-hop link is not worse than the statistical delay performance specified by (ε, D_{\max}) , where ε is the limitation on the statistical delay violation probability, and D_{\max} is the maximum tolerable delay. Note that the end-to-end delay consists of the queueing and transmission delays. As indicated in [26, Section IV], the flow of data bits are treated as the flow of a fluid in the theory of effective bandwidth, in which case the transmission delay can be negligible if $T \ll D_{\max}$. The end-to-end delay can be approximated by the queueing end-to-end delay [7], [8]. Assume that the first-in first-out (FIFO) queues are saturated, and hence they always attempt to transmit [25]. Then, the queueing delay violation probability can be written equivalently as [7], [8]

$$\Pr\{D > D_{\max}\} \doteq e^{-J(\theta)D_{\max}} \quad (5)$$

where we define $f(x) \doteq e^{cx}$ when $\lim_{x \rightarrow \infty} \frac{\log f(x)}{x} = c$, and

$$J(\theta) = \theta\delta = -\Lambda_C(-\theta) \quad (6)$$

is the statistical delay exponent associated with the queue, with $\Lambda_C(\theta)$ denoting the LMGF of the service rate, and δ is decided by the arrival and departure processes jointly. Note that the larger $J(\theta)$, the smaller the delay violation probability is, implying more stringent delay constraints. Now, we can express the probability density function of the random variable D as

$$p_D(x) = \frac{\partial}{\partial x} (1 - \Pr\{D > x\}) \doteq J(\theta)e^{-J(\theta)x}. \quad (7)$$

Consider the two concatenated queues as depicted in Fig. 1. For the queueing constraints specified

by θ_1 and θ_2 with (3) satisfied for each queue, we define

$$J_1(\theta_1) = -\Lambda_{C,1}(-\theta_1), \text{ and } J_2(\theta_2) = -\Lambda_{C,2}(-\theta_2), \quad (8)$$

where $\Lambda_{C,1}(\theta_1)$ and $\Lambda_{C,2}(\theta_1)$ are the LMGFs of the service rates of queues at the source and relay nodes, respectively. In the two-hop system, we can express the end-to-end delay violation probability as

$$\Pr\{D_1 + D_2 > D_{\max}\} = 1 - \int_0^{D_{\max}} \int_0^{D_{\max}-D_1} p_D(D_1) p_D(D_2) dD_2 dD_1 \quad (9)$$

$$\doteq \begin{cases} \frac{J_1(\theta_1)e^{-J_2(\theta_2)D_{\max}} - J_2(\theta_2)e^{-J_1(\theta_1)D_{\max}}}{J_1(\theta_1) - J_2(\theta_2)}, & J_1(\theta_1) \neq J_2(\theta_2), \\ (1 + J_1(\theta_1)D_{\max})e^{-J_1(\theta_1)D_{\max}}, & J_1(\theta_1) = J_2(\theta_2). \end{cases} \quad (10)$$

Note that we should satisfy

$$\Pr\{D_1 + D_2 > D_{\max}\} \leq \varepsilon. \quad (11)$$

C. Effective Capacity

We can dynamically control the delay constraints at the queues of the source and relay nodes specified by $J_1(\theta_1)$ and $J_2(\theta_2)$ as long as the statistical end-to-end delay performance (11) can be guaranteed. At the same time, for each realization of (θ_1, θ_2) , assume that the constant arrival rate at the source is $R \geq 0$, and the channels operate at their capacities. To satisfy the queueing constraint at the source, we must have

$$\tilde{\theta} \geq \theta_1, \quad (12)$$

where $\tilde{\theta}$ is the solution to

$$R = -\frac{\Lambda_{sr}(-\tilde{\theta})}{\tilde{\theta}}, \quad (13)$$

and $\Lambda_{sr}(\theta)$ is the LMGF of the instantaneous capacity of the **S** – **R** link.

In order to satisfy the queueing constraint of the intermediate relay node **R**, we must have

$$\hat{\theta} \geq \theta_2, \quad (14)$$

where $\hat{\theta}$ is the solution to

$$\Lambda_r(\hat{\theta}) + \Lambda_{rd}(-\hat{\theta}) = 0. \quad (15)$$

Above, $\Lambda_r(\theta)$ is the LMGF of the arrival process to the queue at the relay, and $\Lambda_{rd}(\theta)$ is the LMGF of the instantaneous capacity of the **R** – **D** link.

Note that we can obtain the effective capacity $R_E(\theta_1, \theta_2)$ with (θ_1, θ_2) following the method provided in [15, Theorem 2] (Appendix A).¹ Denote Ω as the set of pairs (θ_1, θ_2) such that (11) can be satisfied. After these characterizations, effective capacity of the two-hop communication model under statistical delay constraints (ε, D_{\max}) can be formulated as follows.

Definition 1: The effective capacity of the two-hop communication link with statistical delay constraints specified by (ε, D_{\max}) is given by

$$R_\varepsilon(\varepsilon, D_{\max}) = \sup_{(\theta_1, \theta_2) \in \Omega} R_E(\theta_1, \theta_2) \quad (16)$$

where Ω is the set of all feasible (θ_1, θ_2) satisfying (11). Hence, effective capacity is now the maximum constant arrival rate that can be supported by the two-hop channels under the end-to-end statistical delay constraints.

III. STATISTICAL DELAY TRADEOFFS

For the following analysis, we first characterize the relation between $J_1(\theta_1)$ and the associated minimum $J_2(\theta_2)$ satisfying the statistical delay constraint (11). We have the following results.

Lemma 1: Consider the following function

$$\vartheta(J_1(\theta_1), J_2(\theta_2)) = \frac{J_2(\theta_2)e^{-J_1(\theta_1)D_{\max}} - J_1(\theta_1)e^{-J_2(\theta_2)D_{\max}}}{J_2(\theta_2) - J_1(\theta_1)} = e^{-J_0 D_{\max}} = \varepsilon, \text{ for } 0 \leq \varepsilon \leq 1, \quad (17)$$

where $J_0 = -\frac{\log \varepsilon}{D_{\max}}$ is defined as the statistical delay exponent associated with (ε, D_{\max}) . Denoting $J_2(\theta_2) = \Phi(J_1(\theta_1))$ as a function of $J_1(\theta_1)$, we have the following properties:

a) $\Phi(J_1(\theta_1))$ is continuous. Moreover, for $J_1(\theta_1) = J_{th}(\varepsilon)$, we have

$$\Phi(J_1(\theta_1)) = J_{th}(\varepsilon) \quad (18)$$

where

$$J_{th}(\varepsilon) = -\frac{1}{D_{\max}} \left(1 + \mathcal{W}_{-1} \left(-\frac{\varepsilon}{e} \right) \right), \quad (19)$$

with $\mathcal{W}_{-1}(\cdot)$ denoting the Lambert W function, which is the inverse function of $y = xe^x$ in the range $(-\infty, -1]$.

¹We include the theorem in Appendix A for the reader's convenience.

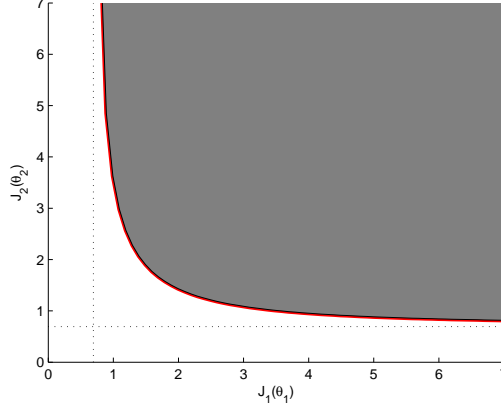


Fig. 2. J_2 v.s. J_1 . $D_{\max} = 1$ sec. $\varepsilon = 0.05$.

- b) Φ is strictly decreasing in $J_1(\theta_1)$.
- c) Φ is convex in $J_1(\theta_1)$.
- d) $J_1(\theta_1) \in [J_0, \infty)$, and $J_2(\theta_2) = \Phi(J_1(\theta_1)) \in [J_0, \infty)$.

Proof: See Appendix B.

Remark 1: The above properties can be understood intuitively. Larger $J_1(\theta_1)$ enforces more stringent delay constraints on queue 1 (i.e., the source queue), and we can loosen the delay constraints for the queue 2 (i.e., the relay queue), and vice versa. When either queue is subject to a deterministic constraint, i.e., $\theta = \infty$, the delay violation occurs only at the other queue. In Fig. 2, we plot J_2 as a function of J_1 for the case with $\varepsilon = 0.05$ and $D_{\max} = 1$ sec for illustration. Note that only (J_1, J_2) in the dark region are feasible to achieve the statistical delay performance. As can be seen from the figure, the curve given by the lower boundary matches the properties in the lemma.

IV. EFFECTIVE CAPACITY IN BLOCK-FADING CHANNELS

In this section, we seek to identify the constant arrival rates R that can be supported by the two-hop system according to the statistical delay tradeoff characterized earlier. We consider a block fading scenario in which the fading stays constant for a block of T seconds and changes independently from one block to another.

We assume that the channel state information (CSI) of the link $\mathbf{S} - \mathbf{R}$ is available at \mathbf{S} and \mathbf{R} , and the CSI of the link $\mathbf{R} - \mathbf{D}$ is available at \mathbf{R} and \mathbf{D} . The instantaneous capacities of the $\mathbf{S} - \mathbf{R}$

and $\mathbf{R} - \mathbf{D}$ links in each block are given, respectively, by

$$C_1 = TB \log_2(1 + \text{SNR}_1 z_1), \quad \text{and} \quad C_2 = TB \log_2(1 + \text{SNR}_2 z_2), \quad (20)$$

in the units of bits per block or equivalently bits per T seconds. These can be regarded as the service processes at the source and relay.

A. Buffer Stability and Log-Moment Generating Function of Block Fading Channels

To ensure the stability of the queues, we need to enforce the following condition [5]

$$\mathbb{E}_{z_1}\{C_1\} < \mathbb{E}_{z_2}\{C_2\}, \quad (21)$$

i.e., the average arrival rate for the queue at the relay should be less than the average service rate.

Under the block fading assumption, the LMGFs for the service processes of queues at the source \mathbf{S} and the relay \mathbf{R} as functions of θ are given by

$$\Lambda_{sr}(\theta) = \log \mathbb{E}_{z_1}\{e^{\theta C_1}\}, \quad \text{and} \quad \Lambda_{rd}(\theta) = \log \mathbb{E}_{z_2}\{e^{\theta C_2}\}. \quad (22)$$

The LMGF for the arrival process of the queue at the relay is [15]

$$\Lambda_r(\theta) = \begin{cases} R\theta, & 0 \leq \theta \leq \tilde{\theta}, \\ R\theta + \log \mathbb{E}_{z_1}\{e^{(\theta - \tilde{\theta})C_1}\}, & \theta > \tilde{\theta}. \end{cases} \quad (23)$$

B. Effective Capacity under Statistical Delay Constraints

In the following, we first assume that there exist θ_1 and θ_2 such that (11) is satisfied. We can identify the effective capacity associated with the given θ_1 and θ_2 values from Theorem 2. Reminding the statistical delay tradeoff indicated in Lemma 1, we can obtain the maximum effective capacity by looping over all possible $(J_1(\theta_1), J_2(\theta_2))$, i.e., θ_1 and θ_2 , which is the effective capacity under the statistical delay constraint in Definition 1.

From (8) and (22), we have

$$J_1(\theta) = -\log \mathbb{E}_{z_1}\{e^{-\theta C_1}\}, \quad \text{and} \quad J_2(\theta) = -\log \mathbb{E}_{z_2}\{e^{-\theta C_2}\}. \quad (24)$$

We can show the following properties of $J(\theta)$.

Lemma 2: Consider the function

$$J(\theta) = -\log \mathbb{E}_z\{e^{-\theta C}\} \quad \text{for} \quad \theta \geq 0, \quad (25)$$

where $C = TB \log_2(1 + \text{SNR}z)$. This function has the following properties.

- a) $J(0) = 0$.
- b) $J(\theta)$ is increasing in θ , and $\dot{J}(0) = \mathbb{E}_z\{C\} > 0$, i.e., the first derivative of $J(\theta)$ with respect to θ at $\theta = 0$ is given by the average service rate.
- c) $J(\theta)$ is a concave function of θ .
- d) $\lim_{\theta \rightarrow \infty} J(\theta) = -\log \Pr\{C = 0\}$, i.e., the negative of the logarithm of the probability of the event that the service rate is 0.

Proof: See Appendix C.

Remark 2: From the properties above, we can see that $J(\theta)$ is equal to 0 at $\theta = 0$, and then it increases sublinearly, and approaches an upperbound, if it exists, as $\theta \rightarrow \infty$. Therefore, $J(\theta)$ is a bijective function of θ , and for each value of J , we can find the associated θ . Note that the effective capacity expressed as $\frac{J(\theta)}{\theta}$ is decreasing in θ [15].

Remark 3: In the remainder of the paper, we use the following definitions

$$R_1 = \frac{J_1(\theta_1)}{\theta_1}, \quad \text{and} \quad R_2 = \frac{J_2(\theta_2)}{\theta_2}. \quad (26)$$

Assumption 1: Throughout this paper, we consider the fading distributions that satisfy the following conditions: 1) $\Pr\{z_1 = 0\} = 0$; 2) $\Pr\{z_2 = 0\} = 0$.

Remark 4: Under the above assumption, we can see that $J_1(\theta)$ and $J_2(\theta)$ approaches ∞ as θ increases. Note that for the continuous distributions of the fading states, such as Rayleigh and Rician fading, the above assumption is justified immediately. If the above assumption does not hold, we can see that the upper bounds for $J_1(\theta_1)$ and $J_2(\theta_2)$ are finite-valued, and the following analysis still holds while only considering a sliced part of (J_1, J_2) of the $J_1 - J_2$ curve characterized in Lemma 1.

Remark 5: According to Lemma 2 and the conditions specified in (12) and (14), we can see that the effective capacity obtained always satisfies the statistical delay constraints as long as θ_1 and θ_2 satisfy (11). Therefore, with the definitions of $J_1(\theta_1)$ and $J_2(\theta_2)$ in (24), we can find the associated θ_1 and θ_2 on the lower boundary curve indicated by Lemma 1. Iterating over this set of θ_1 and θ_2 , we can derive the maximum effective capacity under end-to-end statistical delay constraints. For other values of θ_1 and θ_2 , either (11) cannot be satisfied, or one of the queues is subject to a more stringent constraint than necessary, decreasing the achievable throughput.

For the following analysis, we define

$$\Omega_\varepsilon = \{(\theta_1, \theta_2) : J_1(\theta_1) \text{ and } J_2(\theta_2) \text{ are solutions to (17)}\}. \quad (27)$$

We can characterize the effective capacity of the two-hop system given the statistical queueing constraints θ_1 and θ_2 in Theorem 2. Now, we are seeking to identify the effective capacity of the two-hop system under statistical delay constraints specified by (ε, D_{\max}) , in which case θ_1 and θ_2 are unknown. Combining the behavior of $R_E(\theta_1, \theta_2)$ given (θ_1, θ_2) and the tradeoff between $J_1(\theta_1)$ and $J_2(\theta_2)$ in Lemma 1, we have the following result. Note that $z_{i,\min}$ and $z_{i,\max}$ denote the minimum and maximum value of z_i , respectively.

Theorem 1: The effective capacity of the two-hop wireless communication system subject to end-to-end statistical delay constraints specified by (ε, D_{\max}) is given by the following:

Case I: If $\theta_{1,th} = \theta_{2,th}$,

$$R_\varepsilon(\varepsilon, D_{\max}) = \frac{J_{th}(\varepsilon)}{\theta_{1,th}}, \quad (28)$$

where $(\theta_{1,th}, \theta_{2,th})$ is the unique solution pair to $J_1(\theta_1) = J_{th}(\varepsilon)$, and $J_2(\theta_2) = J_{th}(\varepsilon)$.

Case II: If $\theta_{1,th} > \theta_{2,th}$,

$$R_\varepsilon(\varepsilon, D_{\max}) = \begin{cases} \frac{J_0}{\theta_{1,0}}, & TB \log_2(1 + \text{SNR}_2 z_{2,\min}) \geq TB \log_2(1 + \text{SNR}_1 z_{1,\max}), \\ \frac{J_1(\theta_1)}{\theta_1}, & \text{otherwise.} \end{cases} \quad (29)$$

where $\theta_{1,0}$ is the solution to $J_1(\theta_1) = J_0$, and θ_1° is the smallest value of θ_1 with $(\theta_1, \theta_2) \in \Omega_\varepsilon$ satisfying

$$-\frac{1}{\theta_1} \log \mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\} = -\frac{1}{\theta_1} \left(\log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} + \log \mathbb{E}_{z_1} \{e^{(\theta_2 - \theta_1) C_1}\} \right). \quad (30)$$

Moreover, if $\frac{dJ_2(\theta)}{d\theta}|_{\theta=\underline{\theta}_1} \leq \frac{dJ_1(\theta)}{d\theta}|_{\theta=\underline{\theta}_1}$, where $\underline{\theta}_1$ is the value of θ_1 with $(\theta_1, \theta_2) \in \Omega_\varepsilon$ satisfying

$$\theta_1 = \theta_2, \quad (31)$$

the solution to (30) with $(\theta_1, \theta_2) \in \Omega_\varepsilon$ is unique.

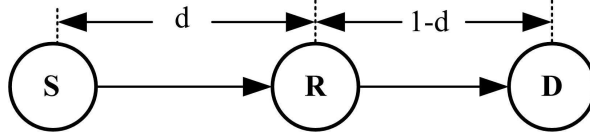


Fig. 3. The relay model.

Case III: If $\theta_{1,th} < \theta_{2,th}$,

$$R_\varepsilon(\varepsilon, D_{\max}) = \begin{cases} \frac{J_0}{\theta_{2,0}}, & TB \log_2(1 + \text{SNR}_1 z_{1,\min}) \geq \frac{J_0}{\theta_{2,0}} \\ \frac{J_2(\check{\theta}_2)}{\check{\theta}_2}, & \text{otherwise.} \end{cases} \quad (32)$$

where $\theta_{2,0}$ is the solution to $J_2(\theta_2) = J_0$, and $(\check{\theta}_1, \check{\theta}_2)$ is the unique solution to

$$\frac{J_1(\theta_1)}{\theta_1} = \frac{J_2(\theta_2)}{\theta_2} \quad (33)$$

with $(\theta_1, \theta_2) \in \Omega_\varepsilon$.

Proof: See Appendix D.

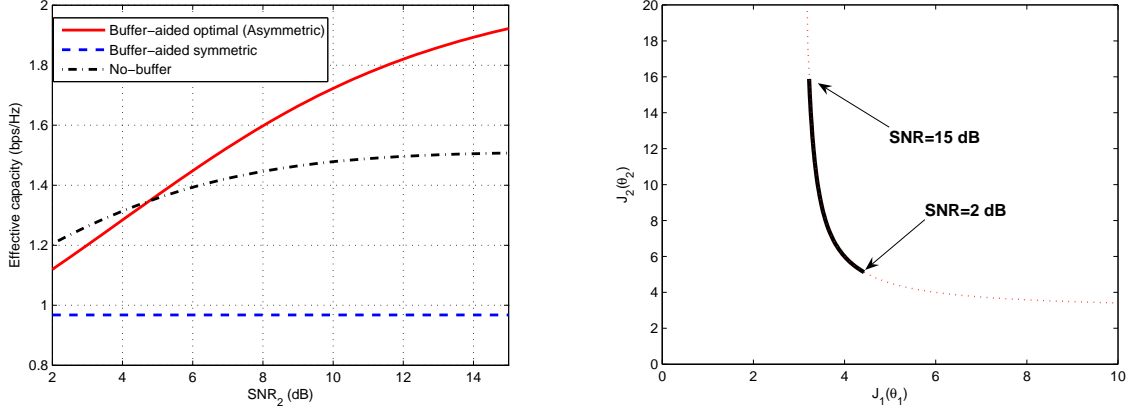
Remark 6: The above theorem covers all the possibilities in which symmetric or asymmetric delay constraints on the queues at the source and relay nodes can be optimal in the sense of achieving the maximum effective capacity of the two-hop relay system. **Case I** refers to the case that the maximum throughput can be achieved with symmetric delay constraints at the queues of the source and relay. **Case II** represents the case when the statistical delay constraints at the relay can be more stringent, while **Case III** shows the scenario with stricter delay constraints at the source. Recalling Theorem 2, we know that as $\varepsilon \rightarrow 1$, $\theta_1 \rightarrow 0$ and $\theta_2 \rightarrow 0$, and hence

$$\lim_{\varepsilon \rightarrow 1} R_\varepsilon(\varepsilon, D_{\max}) = \min \left\{ \lim_{\theta_1 \rightarrow 0} \frac{J_1(\theta_1)}{\theta_1}, \lim_{\theta_2 \rightarrow 0} \frac{J_2(\theta_2)}{\theta_2} \right\} \quad (34)$$

$$= \min \{ \mathbb{E}\{C_1\}, \mathbb{E}\{C_2\} \}. \quad (35)$$

V. NUMERICAL RESULTS

We consider the relay model depicted in Fig. 3. The source, relay, and destination nodes are located on a straight line. The distance between the source and the destination is normalized to 1. Let the distance between the source and the relay node be $d \in (0, 1)$. Then, the distance between the relay and the destination is $1 - d$. We assume the fading distributions for $\mathbf{S} - \mathbf{R}$ and $\mathbf{R} - \mathbf{D}$ links follow independent Rayleigh fading with means $\mathbb{E}\{z_1\} = 1/d^\alpha$ and $\mathbb{E}\{z_2\} = 1/(1-d)^\alpha$, respectively, where



(a) Effective capacity vs. SNR_2 . $\text{SNR}_1 = 0$ dB. $\varepsilon = 0.05$. (b) $J_2(\theta_2)$ vs. $J_1(\theta_1)$ as SNR_2 varies. $\text{SNR}_1 = 0$ dB. $\varepsilon = 0.05$.

Fig. 4. Effective capacity as a function of SNR_2 .

we assume that the path loss $\alpha = 4$. We assume that $D_{\max} = 1$ sec, and $\text{SNR}_1 = 0$ dB in the following numerical results. The curve “Buffer-aided optimal (Asymmetric)” stands for the results in Theorem 1. We also plot the achievable rate when there is no buffer at the relay node “No-buffer” [12], i.e., the service rate of the queue at the source is given by $\frac{TB}{2} \min\{\log_2(1 + 2\text{SNR}_1 z_1), \log_2(1 + 2\text{SNR}_2 z_2)\}$ [27], and the effective capacity with symmetric delay constraints for the two queues “Buffer-aided symmetric”, i.e., $J_1(\theta_1) = J_2(\theta_2) = J_{th}(\varepsilon)$ [18], [19].

In Fig. 4(a), we plot the effective capacity as a function of SNR of the relay node. We fix $d = 0.5$, in which case the $S - R$ and $R - D$ links experience the same channel conditions on average. We assume that the maximum delay violation probability is $\varepsilon = 0.05$. From the figure, we can see that the effective capacity of the two-hop system increases with SNR_2 . Note that at small values of SNR_2 , the buffer at the relay introduces certain loss in the achievable rate. As SNR_2 increases, the buffer at the relay can be beneficial to the two-hop system under statistical delay constraints such that the achievable throughput can be larger. And, in all cases, the achievable rate of asymmetric delay constraints is greater than the one achieved with symmetric delay constraints at the two buffers. In Fig. 4(b), we plot the associated $J_2(\theta_2)$ as a function of $J_1(\theta_1)$. As can be seen from the figure, $J_2(\theta_2)$ increases as SNR_2 increases, i.e., we can impose more stringent constraints to the queue at the relay, and hence the delay constraint at the source can be relaxed. In this way, the effective capacity of the two-hop system can be improved.

We are also interested in the impact of the delay violation probability ε on the achievable per-

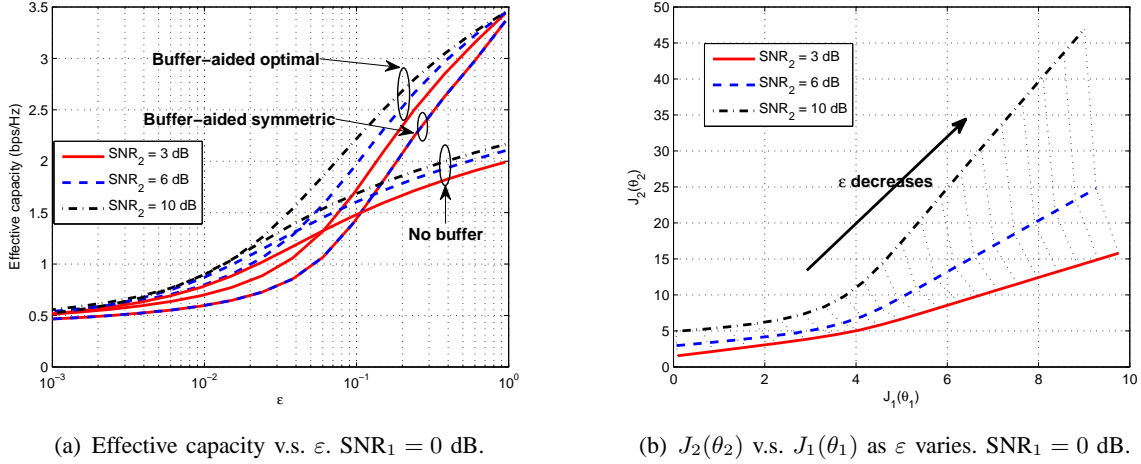


Fig. 5. Effective capacity as a function of ε .

formance. In Fig. 5(a), we plot the effective capacity as ε varies for $\text{SNR}_2 = \{3, 6, 10\}$ dB. It is not surprising that when $\varepsilon \rightarrow 1$, the effective capacities for different SNR_2 are the same, since $R_\varepsilon(\varepsilon, D_{\max}) \rightarrow \min\{\mathbb{E}\{C_1\}, \mathbb{E}\{C_2\}\} = \mathbb{E}\{C_1\}$ in this case. Also, when $\varepsilon \rightarrow 1$, the achievable rate with buffer at the relay is larger than the achievable rate without buffer at the relay, in accordance with the finding in [21] that the throughput can be improved by buffer-aided relay. Moreover, it is interesting that when ε is relatively large but not one, i.e., the statistical delay constraints are less stringent, the achievable throughput with buffer at the relay is larger. Therefore, buffer-aided relay can be helpful even in the presence of end-to-end delay constraints for certain cases. Also, we can find that for larger SNR_2 , the buffer at the relay can help improve the achievable rate at a smaller ε , i.e., in the presence of more stringent delay constraints. To get more insights, we also plot the associated values of $J_1(\theta_1)$ and $J_2(\theta_2)$ as ε decreases in Fig. 5(b). We can see that the increase in $J_2(\theta_2)$ becomes larger in comparison with $J_1(\theta_1)$. Considering the convexity of $J_2(\theta_2)$ in $J_1(\theta_1)$ in Lemma 1, loosening the queueing constraint at one queue will require the other queue to operate in a much more conservative way, which provides little gain under more stringent delay constraints, i.e., for smaller ε .

In Fig. 6, we plot the effective capacity as d varies. We assume $\text{SNR}_2 = \{3, 6, 10\}$ dB, $\varepsilon = 0.05$. We can see from the figure that as d increases, i.e., the channel condition at the link $\mathbf{S} - \mathbf{R}$ is worse, the effective capacity decreases, and the increase of SNR at the relay node helps little. It is interesting that even for small values of SNR_2 , as d increases, the buffer at the relay can help improve the achievable throughput. Albeit, the benefits provided by the buffer at the relay vanish as d approaches 1 since

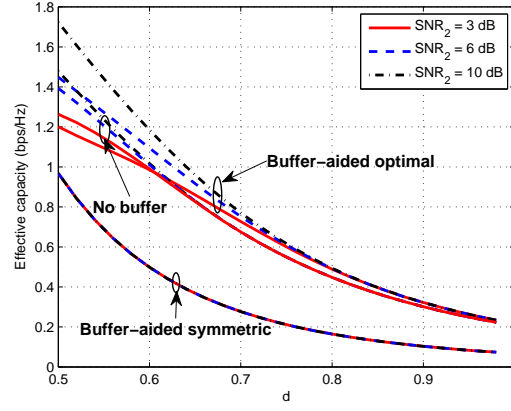
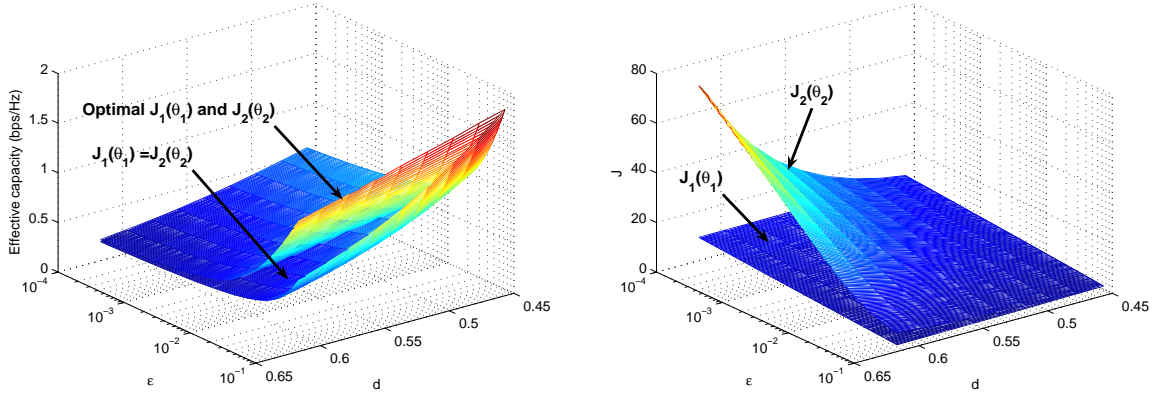


Fig. 6. Effective capacity as a function of d . $\text{SNR}_1 = 0$ dB. $\varepsilon = 0.05$.



(a) Effective capacity v.s. d and ε . $\text{SNR}_1 = 0$ dB. $\text{SNR}_2 = 3$ dB. (b) $J_1(\theta_1)$ and $J_2(\theta_2)$ as functions of d and ε . $\text{SNR}_1 = 0$ dB. $\text{SNR}_2 = 3$ dB.

Fig. 7. Effective capacity as a function of d and ε .

the link $\mathbf{S} - \mathbf{R}$ becomes the bottleneck of the system. Finally, we plot the effective capacity as d and ε vary in Fig. 7(a), with the associated delay tradeoff $J_1(\theta_1)$ and $J_2(\theta_2)$ for the proposed asymmetric delay constraints in Fig. 7(b). We assume $\text{SNR}_2 = 3$ dB. As can be seen from the figure, for all cases, effective capacity decreases as d increases or ε decreases. The improvement in effective capacity is achieved through strong bias towards the queue at the source, in which case we have much larger $J_2(\theta_2)$ in comparison with $J_1(\theta_1)$.

VI. CONCLUSION

In this paper, we have investigated the maximum constant arrival rates that can be supported by a two-hop communication link with a buffer-aided relay under end-to-end statistical delay constraints. We have provided a unified framework for achieving the statistical delay tradeoffs imposed to the source and relay nodes while satisfying the statistical delay constraints. We have determined the effective capacity in the block-fading scenario as a function of the statistical delay constraints, the signal-to-noise ratio levels SNR_1 and SNR_2 , and the fading distributions. We have shown that asymmetric delay constraints at the two buffers can help increase the effective capacity of the two-hop system compared with symmetric delay constraints. We have found that buffer-aided relay can improve the achievable rate of the system under delay constraints when the SNR at the relay is high, the end-to-end delay constraints is loose, or when the channel conditions between the relay and destination node are more favorable.

APPENDIX

A. Preliminary Results

Proposition 1: ([15]) The constant arrival rates, which can be supported by the two-hop link in the presence of queueing constraints θ_1 and θ_2 at the source and relay, respectively, are upperbounded by

$$R \leq \min \left\{ -\frac{1}{\theta_1} \log \mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\}, -\frac{1}{\theta_2} \log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} \right\} = \min \left\{ \frac{J_1(\theta_1)}{\theta_1}, \frac{J_2(\theta_2)}{\theta_2} \right\}. \quad (36)$$

Theorem 2: ([15]) The effective capacity of the two-hop system given $\theta_1 > 0$ and $\theta_2 > 0$ is given by the following:

Case I: If $\theta_1 \geq \theta_2$,

$$R_E(\theta_1, \theta_2) = \min \left\{ -\frac{1}{\theta_1} \log \mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\}, -\frac{1}{\theta_2} \log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} \right\}. \quad (37)$$

Case II: If $\theta_1 < \theta_2$ and $\theta_2 \leq \bar{\theta}$,

$$R_E(\theta_1, \theta_2) = -\frac{1}{\theta_1} \log \mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\} \quad (38)$$

where $\bar{\theta}$ is the unique value of θ for which we have the following equality satisfied:

$$-\frac{1}{\theta_1} \log \mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\} = -\frac{1}{\theta_1} \left(\log \mathbb{E}_{z_2} \{e^{-\theta C_2}\} + \log \mathbb{E}_{z_1} \{e^{(\theta-\theta_1)C_1}\} \right). \quad (39)$$

Case III: Assume $\theta_1 < \theta_2$ and $\theta_2 > \bar{\theta}$.

III.a: If

$$-\frac{1}{\theta_2} \log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} \geq -\frac{1}{\theta_2} \log \mathbb{E}_{z_1} \{e^{-\theta_2 C_1}\}, \quad (40)$$

then

$$R_E(\theta_1, \theta_2) = -\frac{1}{\tilde{\theta}^*} \log \mathbb{E}_{z_1} \{e^{-\tilde{\theta}^* C_1}\} \quad (41)$$

where $\tilde{\theta}^*$ is the smallest solution to

$$-\frac{1}{\tilde{\theta}} \log \mathbb{E}_{z_1} \{e^{-\tilde{\theta} C_1}\} = -\frac{1}{\tilde{\theta}} \left(\log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} + \log \mathbb{E}_{z_1} \{e^{(\theta_2 - \tilde{\theta}) C_1}\} \right). \quad (42)$$

III.b: Otherwise,

$$R_E(\theta_1, \theta_2) = -\frac{1}{\theta_2} \log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\}. \quad (43)$$

B. Proof of Lemma 1

1) When $J_1(\theta_1) \neq J_2(\theta_2)$, the continuity is obvious since there is no pole to (17). Consider $J_1(\theta_1) = J_2(\theta_2)$. We can see that

$$\lim_{J_2(\theta_2) \rightarrow J_1(\theta_1)_-} \vartheta(J_1(\theta_2), J_2(\theta_2)) = \lim_{J_2(\theta_2) \rightarrow J_1(\theta_1)_-} \frac{J_2(\theta_2)e^{-J_1(\theta_1)D_{\max}} - J_1(\theta_1)e^{-J_2(\theta_2)D_{\max}}}{J_2(\theta_2) - J_1(\theta_1)} \quad (44)$$

$$= \lim_{J_2(\theta_2) \rightarrow J_1(\theta_1)_-} e^{-J_2(\theta_2)D_{\max}} \frac{J_2(\theta_2)e^{-(J_1(\theta_1) - J_2(\theta_2))D_{\max}} - J_1(\theta_1)}{J_2(\theta_2) - J_1(\theta_1)} \quad (45)$$

$$= \lim_{J_2(\theta_2) \rightarrow J_1(\theta_1)_-} e^{-J_2(\theta_2)D_{\max}} \left(1 + J_2(\theta_2) \frac{1 - e^{-(J_1(\theta_1) - J_2(\theta_2))D_{\max}}}{J_1(\theta_1) - J_2(\theta_2)} \right) \quad (46)$$

$$= e^{-J_2(\theta_2)D_{\max}} (1 + J_2(\theta_2)D_{\max}). \quad (47)$$

Similarly, we can show that

$$\lim_{J_2(\theta_2) \rightarrow J_1(\theta_1)_+} \vartheta(J_1(\theta_2), J_2(\theta_2)) = e^{-J_1(\theta_1)D_{\max}} (1 + J_1(\theta_1)D_{\max}). \quad (48)$$

From (10), we can see that at $J_1(\theta_1) = J_2(\theta_2)$, $\vartheta(J_1(\theta_2), J_2(\theta_2))$ is continuous, i.e., $J_2(\theta_2) = \Phi(J_1(\theta_1))$ is continuous, and from (11), we should have

$$(1 + J_1(\theta_1)D_{\max}) e^{-J_1(\theta_1)D_{\max}} \leq \varepsilon \quad (49)$$

which gives us (19) immediately by solving the above equation with equality.

- 2) Taking the partial derivative of $\vartheta(J_1(\theta_1), J_2(\theta_2))$ in $J_1(\theta_1)$ and noting that the right-hand-side (RHS) of (17) is constant, we have

$$\begin{aligned} \frac{\partial \vartheta(J_1(\theta_1), J_2(\theta_2))}{\partial J_1(\theta)} &= \frac{1}{(J_2(\theta_2) - J_1(\theta_1))^2} \left(\left(j_2(\theta) e^{-J_1(\theta_1) D_{\max}} - J_2(\theta_2) D_{\max} e^{-J_1(\theta_1) D_{\max}} - e^{-J_2(\theta_2) D_{\max}} \right. \right. \\ &\quad \left. \left. + J_1(\theta) j_2(\theta_2) D_{\max} e^{-J_2(\theta_2) D_{\max}} \right) (J_2(\theta_2) - J_1(\theta_1)) - (\dot{J}_2(\theta_2) - 1) \right. \\ &\quad \left. \times \left(J_2(\theta_2) e^{-J_1(\theta_1) D_{\max}} - J_1(\theta_1) e^{-J_2(\theta_2) D_{\max}} \right) \right) = 0, \end{aligned} \quad (50)$$

which, after combining the coefficients of $\dot{J}_2(\theta_2)$ and rearrangements, gives us

$$\dot{\Phi}(J_1(\theta_1)) = \dot{J}_2(\theta_2) = \frac{J_2(\theta_2)}{J_1(\theta_1)} e^{(J_2(\theta_2) - J_1(\theta_1)) D_{\max}} \frac{(J_2(\theta_2) - J_1(\theta_1)) D_{\max} + e^{-(J_2(\theta_2) - J_1(\theta_1)) D_{\max}} - 1}{(J_2(\theta_2) - J_1(\theta_1)) D_{\max} + 1 - e^{(J_2(\theta_2) - J_1(\theta_1)) D_{\max}}} \quad (51)$$

In the following, we will show that $\dot{\Phi}(J_1(\theta_1)) < 0$. Denote $x = (J_2(\theta_2) - J_1(\theta_1)) D_{\max}$, and define

$$\nu(x) = \frac{x + e^{-x} - 1}{x + 1 - e^x}. \quad (52)$$

Then, we can rewrite $\dot{\Phi}(J_1(\theta))$ as

$$\dot{\Phi}(J_1(\theta_1)) = \dot{J}_2(\theta_2) = \frac{J_2(\theta_2)}{J_1(\theta_1)} e^{(J_2(\theta_2) - J_1(\theta_1)) D_{\max}} \nu(x). \quad (53)$$

Note that $\frac{J_2(\theta_2)}{J_1(\theta_1)} e^{(J_2(\theta_2) - J_1(\theta_1)) D_{\max}}$ is positive. Taking the first derivative of $\nu(x)$, we obtain

$$\dot{\nu}(x) = \frac{4 - 2(e^x + e^{-x}) + x(e^x - e^{-x})}{(x + 1 - e^x)^2} \quad (54)$$

We can show that $\dot{\nu}(x) \geq 0$. Suppose $x > 0$. Considering the numerator of the above equation, we have

$$4 - 2(e^x + e^{-x}) + x(e^x - e^{-x}) = -2(e^{\frac{x}{2}} - e^{-\frac{x}{2}})^2 + x(e^{\frac{x}{2}} - e^{-\frac{x}{2}})(e^{\frac{x}{2}} + e^{-\frac{x}{2}}) \quad (55)$$

$$= (e^{\frac{x}{2}} - e^{-\frac{x}{2}})(-2(e^{\frac{x}{2}} - e^{-\frac{x}{2}}) + x(e^{\frac{x}{2}} + e^{-\frac{x}{2}})) \quad (56)$$

$$= e^{-\frac{x}{2}}(x + 2)(e^{\frac{x}{2}} - e^{-\frac{x}{2}}) \left(\frac{x - 2}{x + 2} e^x + 1 \right) \quad (57)$$

$$\geq 0 \quad (58)$$

where $\frac{x-2}{x+2} e^x \geq -1$ is incorporated since it is an increasing function of x , and its value at $x = 0$ is -1 . Therefore, $\dot{\nu}(x) > 0$ for $x > 0$, i.e., $\nu(x)$ is increasing for $x > 0$. In a similar way, we can show that $\dot{\nu}(x) > 0$ for $x < 0$. Additionally, we can show $\lim_{x \rightarrow 0} \dot{\nu}(x) = 0$ by considering the Taylor expansions of e^x and e^{-x} at $x = 0$ and noting that the numerator goes to 0 in the

order $o(x^4)$ while the denominator goes to 0 in the order of x^4 . Therefore, ν is increasing in x . Meanwhile,

$$\lim_{x \rightarrow \infty} \nu(x) = \lim_{x \rightarrow \infty} \frac{x + e^{-x} - 1}{x + 1 - e^x} = \lim_{x \rightarrow \infty} \frac{1 - e^{-x}}{1 - e^x} = 0. \quad (59)$$

Hence, $\nu(x) < 0$, which in turn, tells us that $\dot{\Phi}(J_1(\theta_1)) < 0$ in (53). Therefore, $J_2(\theta_2) = \Phi(J_1(\theta_1))$ is strictly decreasing in $J_1(\theta)$.

- 3) We will show the convexity of Φ by considering the branches for $J_2(\theta_2) > J_1(\theta_1)$ and $J_2(\theta_2) < J_1(\theta_1)$, respectively.

For $J_1(\theta_1) < J_{th}(\varepsilon)$, we know that $J_2(\theta_2) > J_1(\theta_1)$. Consider

$$\dot{J}_2(\theta_2) = \frac{J_2(\theta_2)}{J_1(\theta_1)} e^{(J_2(\theta_2) - J_1(\theta_1))D_{\max}} \frac{(J_2(\theta_2) - J_1(\theta_1))D_{\max} + e^{-(J_2(\theta_2) - J_1(\theta_1))D_{\max}} - 1}{(J_2(\theta_2) - J_1(\theta_1))D_{\max} + 1 - e^{(J_2(\theta_2) - J_1(\theta_1))D_{\max}}} \quad (60)$$

$$= \frac{J_2(\theta_2)}{J_1(\theta_1)} e^x \nu(x) \quad (61)$$

where again $x = (J_2(\theta_2) - J_1(\theta_1))D_{\max}$. Note that as x increases, $\frac{J_2(\theta_2)}{J_1(\theta_1)}$ should increase since $J_1(\theta_1)$ decreases and $J_2(\theta_2)$ increases. From the above discussion, we know $\nu(x) < 0$, for $x > 0$. Define $\eta(x) = e^x \nu(x)$, $\eta(x) < 0$ for $x > 0$. Then, if we can show that $\eta(x)$ is decreasing as x increases, then $\dot{J}_2(\theta_2) = \dot{\Phi}(J_1(\theta_1))$ will decrease with x , since a smaller negative value multiplied with a larger positive value will lead to a smaller negative value. Taking the first derivative of $\eta(x)$, we have

$$\dot{\eta}(x) = e^x(\nu(x) + \dot{\nu}(x)) = e^x \frac{2 + x^2 - (e^x + e^{-x})}{(x + 1 - e^x)^2}. \quad (62)$$

Note that the numerator $2 + x^2 - (e^x + e^{-x})$ can be shown to be less than 0 for $x > 0$. More specifically, consider that its second derivative $2 - (e^x + e^{-x})$ is less than 0 for $x > 0$ and the first derivative $2x - (e^x - e^{-x})$ at $x = 0$ is 0, and hence its first derivative is always less than 0, which tells us that it is a decreasing function in x with the maximum value at $x = 0$ as 0. Therefore, $\dot{\eta} < 0$. Hence, $\dot{J}_2(\theta_2) < 0$ is decreasing as $J_1(\theta_1)$ decreases for $J_1(\theta_1) < J_{th}(\varepsilon)$, i.e., $\ddot{\Phi}(J_1(\theta)) \geq 0$. Similarly, we can show that $\ddot{\Phi}(J_1(\theta_1)) \geq 0$ for $J_1(\theta_1) > J_{th}(\varepsilon)$. Together, we know that $\ddot{\Phi}(J_1(\theta_1)) \geq 0$, and hence $J_2(\theta_2) = \Phi(J_1(\theta_1))$ is a convex function in $J_1(\theta_1)$.

- 4) Letting $J_1(\theta)$ go to infinity, we can see that

$$\lim_{J_1(\theta) \rightarrow \infty} \vartheta(J_1(\theta), J_2(\theta)) = \lim_{J_1(\theta) \rightarrow \infty} e^{-J_2(\theta_2)D_{\max}} = e^{-J_0 D_{\max}} \quad (63)$$

which indicates $\lim_{J_1(\theta) \rightarrow \infty} J_2(\theta_2) = J_0$. On the other hand, if we let $J_2(\theta)$ go to infinity, we

can show that $\lim_{J_2(\theta) \rightarrow \infty} J_1(\theta_1) = J_0$. Together, we obtain the result in the lemma. \square

C. Proof of Lemma 2

- a) This property can be readily seen by evaluating the function at $\theta = 0$.
- b) The first derivative of J with respect to θ can be evaluated as

$$\dot{J}(\theta) = \frac{\mathbb{E}_z \{e^{-\theta C} C\}}{\mathbb{E}_z \{e^{-\theta C}\}} > 0. \quad (64)$$

Then, $\dot{J}(0)$ can be obtained by evaluating the above equation at $\theta = 0$.

- c) The second derivative of J with respect to θ can be expressed as

$$\ddot{J}(\theta) = -\frac{1}{(\mathbb{E}_z \{e^{-\theta C}\})^2} \left(\mathbb{E}_z \{e^{-\theta C} C^2\} \mathbb{E}_z \{e^{-\theta C}\} - (\mathbb{E}_z \{e^{-\theta C} C\})^2 \right). \quad (65)$$

By the Cauchy-Schwarz inequality, we know that $\mathbb{E}\{X^2\}\mathbb{E}\{Y^2\} \geq (\mathbb{E}\{XY\})^2$. Then, denoting $X = \sqrt{e^{-\theta C} C^2}$ and $Y = \sqrt{e^{-\theta C}}$, we easily see that $\ddot{J}(\theta) \leq 0$ for all θ . Thus, $J(\theta)$ is a concave function.

- d) Note that as long as $C \neq 0$, $\lim_{\theta \rightarrow \infty} e^{-\theta C} = 0$, and whenever $C = 0$, $e^{\theta C} = 1$. Therefore, we have $\lim_{\theta \rightarrow \infty} \mathbb{E}_{z \neq 0} \{e^{-\theta C}\} = 0$. Then $\lim_{\theta \rightarrow \infty} J(\theta) = \lim_{\theta \rightarrow \infty} -\log (\mathbb{E}_{z \neq 0} \{e^{-\theta C}\} + \mathbb{E}_{z=0} \{1\}) = -\log \Pr\{C = 0\}$. \square

D. Proof of Theorem 1

With the delay tradeoff specified in Lemma 1, we can see that there is potential improvement of effective capacity by adjusting the statistical delay constraint imposed on the queues at the source and relay nodes. As a starting point, we consider $J_1(\theta_1) = J_2(\theta_2)$. According to Lemma 2 and the subsequent discussions, we can always find $\theta_{1,th}$ and $\theta_{2,th}$ for $J_{th}(\varepsilon)$ defined in (19). Now, depending on the values of $\theta_{1,th}$ and $\theta_{2,th}$, we have different behaviors of the effective capacity in Theorem 2 in Appendix A. We seek to find the optimal $J_1(\theta_1)$ and $J_2(\theta_2)$ with $(\theta_1, \theta_2) \in \Omega_\varepsilon$ to maximize the effective capacity, where Ω_ε is defined in (27).

Case I: Assume $\theta_{1,th} = \theta_{2,th}$. For this case, we should have

$$R_E(\theta_{1,th}, \theta_{2,th}) = R_1 = \frac{J_{th}(\varepsilon)}{\theta_{1,th}} = \frac{J_{th}(\varepsilon)}{\theta_{2,th}} = R_2. \quad (66)$$

We assert that this value is the effective capacity of the two-hop system, i.e.,

$$R_\varepsilon(\varepsilon, D_{\max}) = \sup_{(\theta_1, \theta_2) \in \Omega} R_E(\theta_1, \theta_2) = R_E(\theta_{1,th}, \theta_{2,th}). \quad (67)$$

We can show this by contradiction. We know that the effective capacity is a decreasing function in θ . Suppose that there exists some $R > R_E(\theta_{1,th}, \theta_{2,th})$ that can be supported by the two-hop system with θ_1 and θ_2 . Then, we must have $\theta_1 < \theta_{1,th}$, and hence $J_1(\theta_1) < J_1(\theta_{1,th})$. According to the statistical delay tradeoff shown in Lemma 1, we can see that $J_2(\theta_2) > J_2(\theta_{2,th})$, which tells us that $\theta_2 > \theta_{2,th}$ according to Lemma 2, since $J_2(\theta)$ is increasing in θ . Now, from the Proposition 1 in Appendix A, we obtain

$$R \leq \min \left\{ \frac{J_1(\theta_1)}{\theta_1}, \frac{J_2(\theta_2)}{\theta_2} \right\} = \frac{J_2(\theta_2)}{\theta_2} < \frac{J_2(\theta_{2,th})}{\theta_{2,th}} = R_E(\theta_{1,th}, \theta_{2,th}) \quad (68)$$

which leads to a contradiction.

Case II: Assume $\theta_{1,th} > \theta_{2,th}$. In this case, we can see that

$$R_1 = \frac{J_1(\theta_{1,th})}{\theta_{1,th}} = \frac{J_{th}(\varepsilon)}{\theta_{1,th}} < \frac{J_{th}(\varepsilon)}{\theta_{2,th}} = \frac{J_2(\theta_{2,th})}{\theta_{2,th}} = R_2. \quad (69)$$

The effective capacity associated with $\theta_{1,th}, \theta_{2,th}$ specializes into **Case I** of Theorem 2. Therefore, $R_E(\theta_{1,th}, \theta_{2,th}) = \min\{R_1, R_2\} = R_1$. Obviously, the queueing constraint imposed at the source is more stringent. To achieve better performance, we should try to relax the queueing constraints at the source, i.e., decrease θ_1 , or $J_1(\theta_1)$ equivalently. Correspondingly, from Lemma 1, $J_2(\theta_2)$ should increase, and we have $J_2(\theta_2) > J_{th}(\varepsilon) > J_1(\theta_1)$. In the following, we will provide a characterization of θ_1 as we iterate over $(\theta_1, \theta_2) \in \Omega_\varepsilon$ to find the optimal pair that maximizes the effective capacity.

First, noting that as $J_1(\theta_1)$ decreases from $J_{th}(\varepsilon)$ to J_0 , we can see that θ_1 decreases from $\theta_{1,th}$ to some finite value $\theta_{1,0}$, which is the solution to $J_1(\theta) = J_0$. To the opposite, θ_2 increases from $\theta_{2,th} < \theta_{1,th}$ to ∞ . Clearly, from the continuity of $J_2(\theta_2) = \Phi(J_1(\theta_1))$, the corresponding θ_2 as a function of θ_1 should be continuous as well. Hence, there must be one point $(\underline{\theta}_1, \underline{\theta}_2) \in \Omega_\varepsilon$ such that

$$\underline{\theta}_1 = \underline{\theta}_2, \quad (70)$$

and for all $(\theta_1, \theta_2) \in \Omega_\varepsilon$ with $\theta_1 < \underline{\theta}_1$, we will have $\theta_2 > \underline{\theta}_2 = \underline{\theta}_1 > \theta_1$. According to Lemma 2, we know $J_1(\theta)$ and $J_2(\theta)$ are increasing functions of θ . Therefore, at this point, we have

$$R_1 = \frac{J_1(\underline{\theta}_1)}{\underline{\theta}_1} < \frac{J_1(\theta_{1,th})}{\underline{\theta}_1} = \frac{J_{th}(\varepsilon)}{\underline{\theta}_1} = \frac{J_2(\theta_{2,th})}{\underline{\theta}_1} < \frac{J_2(\underline{\theta}_2)}{\underline{\theta}_1} = \frac{J_2(\underline{\theta}_2)}{\underline{\theta}_2} = R_2. \quad (71)$$

That is, the queue at the source is still the bottleneck of the two-hop system. We can further relieve the queueing constraint at the source.

Now, as θ_1 further decreases, $\theta_1 < \theta_2$. Consequently, the effective capacity associated with $(\theta_1, \theta_2) \in \Omega_\varepsilon$ now specializes into **Case II** of Theorem 2. As can be seen from Theorem 2, the queue at the relay will not affect the performance as long as θ_1 and θ_2 satisfy the following inequality given by

$$-\frac{1}{\theta_1} \log \mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\} \leq -\frac{1}{\theta_1} \left(\log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} + \log \mathbb{E}_{z_1} \{e^{(\theta_2 - \theta_1) C_1}\} \right). \quad (72)$$

Note that as θ_1 decreases from $\underline{\theta}_1$ to $\theta_{1,0}$, the LHS of the above inequality increases from $\frac{J_1(\underline{\theta}_1)}{\underline{\theta}_1}$ to $\frac{J_0}{\theta_{1,0}}$. On the other hand, at $\theta_1 = \underline{\theta}_1$, we have $\underline{\theta}_2 = \underline{\theta}_1$, and the value of the RHS of the above inequality at $(\underline{\theta}_1, \underline{\theta}_2)$ is given by

$$\text{RHS} = \frac{J_2(\underline{\theta}_2)}{\underline{\theta}_1} > \frac{J_1(\underline{\theta}_1)}{\underline{\theta}_1}. \quad (73)$$

As $\theta_1 \rightarrow \theta_{1,0}$, or $J_1(\theta_1) \rightarrow J_0$, we know that

$$\begin{aligned} \lim_{J_1(\theta_1) \rightarrow J_0} \text{RHS} &= \lim_{J_1(\theta_1) \rightarrow J_0} -\frac{1}{\theta_1} \left(\log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} + \log \mathbb{E}_{z_1} \{e^{(\theta_2 - \theta_1) C_1}\} \right) \\ &= \lim_{J_1(\theta_1) \rightarrow J_0} \frac{\theta_2}{\theta_1} \left(-\frac{1}{\theta_2} \log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} - \frac{1}{\theta_2} \log \mathbb{E}_{z_1} \{e^{(\theta_2 - \theta_1) C_2}\} \right). \end{aligned} \quad (74)$$

Note further that $J_2(\theta_2)$, and hence θ_2 , approaches infinity as $J_1(\theta_1) \rightarrow J_0$. The first term inside the parenthesis goes to the minimum rate of the **R – D** link, i.e., $TB \log_2(1 + \text{SNR}_2 z_{2,\min})$, and the second term goes to the largest rate of the link **S – R**, i.e., $TB \log_2(1 + \text{SNR}_1 z_{1,\max})$. So as long as the smallest rate of **R – D** is less than the largest rate of the link **S – R**, the limit in (74) goes to $-\infty$. It is important to note that if the highest rate of **S – R** can be supported by the link **R – D**, i.e.,

$$TB \log_2(1 + \text{SNR}_2 z_{2,\min}) \geq TB \log_2(1 + \text{SNR}_1 z_{1,\max}), \quad (75)$$

then there is no congestion at the relay node at all. In this case, θ_2 can take any value greater than 0, and the only delay caused is the queue at the source. Therefore, the arrival rates are limited by the **S – R** link, and to satisfy the statistical delay constraints, we have

$$R_\varepsilon(\varepsilon, D_{\max}) = \frac{J_0}{\theta_{1,0}}. \quad (76)$$

Now, we consider the case when (75) is not satisfied. In such cases, $\theta_2 \rightarrow \infty$ as $J_2(\theta_2) \rightarrow \infty$.

From the continuity of the functions, we know that there must be some $(\theta_1, \theta_2) \in \Omega_\varepsilon$ such that the above inequality in (72) is satisfied with equality. Denote the smallest θ_1 for such (θ_1, θ_2) pairs as $\overset{\circ}{\theta}_1$. Then, for all $(\theta_1, \theta_2) \in \Omega_\varepsilon$ with $\theta_1 < \overset{\circ}{\theta}_1$, (72) cannot be satisfied.

Moreover, we know as θ_1 decreases, R_1 increases from $\frac{J_{th}(\varepsilon)}{\theta_{1,th}}$ to $\frac{J_0}{\theta_{1,0}}$. At the same time, as θ_2 approaches to infinity, R_2 decreases from $\frac{J_{th}(\varepsilon)}{\theta_{2,th}}$ to $TB \log_2(1 + \text{SNR}_2 z_{\min})$. Therefore, there must be some value such that

$$R_1 = \frac{J_1(\theta_1)}{\theta_1} = R = \frac{J_2(\theta_2)}{\theta_2} = R_2 \quad (77)$$

with the associated statistical queueing constraints denoted as $\check{\theta}_1$ and $\check{\theta}_2$, respectively. For $(\theta_1, \theta_2) \in \Omega_\varepsilon$ with $\theta_1 < \check{\theta}_1$, we have

$$R_1 = \frac{J_1(\theta_1)}{\theta_1} > \frac{J_2(\theta_2)}{\theta_2} = R_2. \quad (78)$$

In the following, we can establish the comparison between $\check{\theta}_1$ and $\overset{\circ}{\theta}_1$ as

$$\check{\theta}_1 \leq \overset{\circ}{\theta}_1. \quad (79)$$

Note here that if $\frac{J_0}{\theta_{1,0}} < TB \log_2(1 + \text{SNR}_2 z_{\min})$, there is no θ_1 for (77) to be satisfied, and hence we can set $\check{\theta}_1$ to be 0 without affecting the following discussion based on $\overset{\circ}{\theta}_1$, which satisfies the above claim obviously. Suppose that $\check{\theta}_1 > \overset{\circ}{\theta}_1$. Since at $\overset{\circ}{\theta}_1$, the condition for **Case II** of Theorem 2 can be satisfied, we immediately see that

$$R_E(\overset{\circ}{\theta}_1, \overset{\circ}{\theta}_2) = \frac{J_1(\overset{\circ}{\theta}_1)}{\overset{\circ}{\theta}_1}. \quad (80)$$

However, according to Proposition 1 and (78), we have

$$R_E(\overset{\circ}{\theta}_1, \overset{\circ}{\theta}_2) \leq \min \left\{ \frac{J_1(\overset{\circ}{\theta}_1)}{\overset{\circ}{\theta}_1}, \frac{J_2(\overset{\circ}{\theta}_2)}{\overset{\circ}{\theta}_2} \right\} = \frac{J_2(\overset{\circ}{\theta}_2)}{\overset{\circ}{\theta}_2} < \frac{J_1(\overset{\circ}{\theta}_1)}{\overset{\circ}{\theta}_1} \quad (81)$$

leading to contradiction. A numerical result provides a visualization of the aforementioned discussions on $\underline{\theta}_1$, $\overset{\circ}{\theta}_1$, and $\check{\theta}_1$. We consider the the delay constraint given by $(\varepsilon, D_{\max}) = (0.05, 1)$ in Rayleigh fading channel. We assume that $\text{SNR}_1 = 0$ dB, $\text{SNR}_2 = 3$ dB, $T = 1$ ms, and $B = 180$ kHz. We obtain $\theta_{1,th} = 0.0178$, and $\theta_{2,th} = 0.011$. Now, as θ_1 decreases while $(\theta_1, \theta_2) \in \Omega_\varepsilon$, we plot the values of θ_1 and θ_2 in Fig. 8(a), the LHS and RHS of (72) in Fig. 8(b), and the values of R_1 and

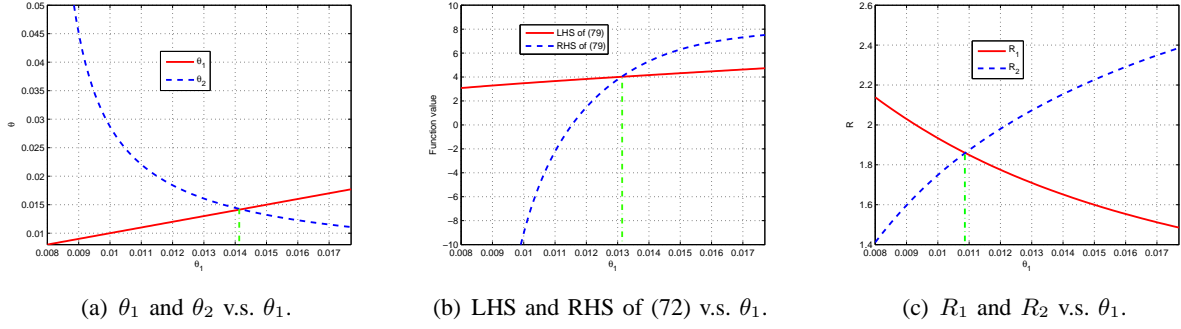


Fig. 8. The illustration of $\underline{\theta}_1$, $\overset{\circ}{\theta}_1$, and $\check{\theta}_1$. From a)-c), the cross points give us $\underline{\theta}_1$, $\overset{\circ}{\theta}_1$, and $\check{\theta}_1$. $\mathbb{E}_{z_1}\{z_1\} = \mathbb{E}_{z_2}\{z_2\} = 16$.

R_2 in Fig. 8(c). We can obtain $\underline{\theta}_1 = 0.0142$, $\overset{\circ}{\theta}_1 = 0.0131$, and $\check{\theta}_1 = 0.0109$. Obviously, we can see that $\check{\theta}_1 < \overset{\circ}{\theta}_1 < \underline{\theta}_1$. Note that we have $\Pr\{z_1 = 0\} = \Pr\{z_2 = 0\} = 0$ for Rayleigh fading channel, and hence $J_1(\theta_2) \rightarrow \infty$ as $\theta_2 \rightarrow \infty$. Note also that $z_{1,\max} = \infty$ and $z_{2,\min} = 0$ for Rayleigh fading channels.

Proposition 2: The effective capacity in this case is given by

$$R_\varepsilon(\varepsilon, D_{\max}) = \sup_{(\theta_1, \theta_2) \in \Omega} R_E(\theta_1, \theta_2) = R_E(\overset{\circ}{\theta}_1, \overset{\circ}{\theta}_2) = \frac{J_1(\overset{\circ}{\theta}_1)}{\overset{\circ}{\theta}_1}. \quad (82)$$

Proof: In order to prove the proposition, we have to show that there is no other arrival rate larger than the value specified above that can be supported by the two-hop link while satisfying the statistical delay constraint. We know that for all $(\theta_1, \theta_2) \in \Omega_\varepsilon$ with $\theta_1 > \overset{\circ}{\theta}_1$,

$$R_E(\theta_1, \theta_2) \leq \frac{J_1(\theta_1)}{\theta_1} < \frac{J_1(\overset{\circ}{\theta}_1)}{\overset{\circ}{\theta}_1} = R_\varepsilon(\varepsilon, D_{\max}). \quad (83)$$

Suppose that there exists $R > R_E(\overset{\circ}{\theta}_1, \overset{\circ}{\theta}_2)$ that can be supported by the two-hop system with θ_1 and θ_2 . Then, $\theta_1 < \overset{\circ}{\theta}_1$. As shown above, for $(\theta_1, \theta_2) \in \Omega_\varepsilon$ with $\theta_1 < \overset{\circ}{\theta}_1$, the inequality defined in (72) cannot be satisfied, and hence $R_E(\theta_1, \theta_2)$ falls into **Case III** of Theorem 2. In addition, with the previous characterization in (70), we know $\theta_2 > \underline{\theta}_2 = \underline{\theta}_1 > \overset{\circ}{\theta}_1$.

For **Case III.b** of Theorem 2, i.e.,

$$\frac{J_2(\theta_2)}{\theta_2} < \frac{J_1(\theta_2)}{\theta_2}, \quad (84)$$

we know that the effective capacity is decreasing in θ , and as a result

$$R_E(\theta_1, \theta_2) = \frac{J_2(\theta_2)}{\theta_2} < \frac{J_1(\theta_2)}{\theta_2} < \frac{J_1(\underline{\theta}_2)}{\underline{\theta}_2} = \frac{J_1(\underline{\theta}_1)}{\underline{\theta}_1} \leq \frac{J_1(\overset{\circ}{\theta}_1)}{\overset{\circ}{\theta}_1} = R_E(\varepsilon, D_{\max}) \quad (85)$$

where $\theta_2 > \underline{\theta}_2 = \underline{\theta}_1 > \overset{\circ}{\theta}_1$ is incorporated.

For **Case III.a** of Theorem 2, there exists $\tilde{\theta}_1^* \in (\theta_1, \theta_2)$ such that $\tilde{\theta}_1^*$ is the smallest solution to

$$-\frac{1}{\tilde{\theta}} \log \mathbb{E}_{z_1} \{e^{-\tilde{\theta} C_1}\} = -\frac{1}{\tilde{\theta}} \left(\log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} + \log \mathbb{E}_{z_1} \{e^{(\theta_2 - \tilde{\theta}) C_1}\} \right).$$

With the assumption $R > R_E(\overset{\circ}{\theta}_1, \overset{\circ}{\theta}_2)$, we must have $\theta_1 < \tilde{\theta}_1^* < \overset{\circ}{\theta}_1$, and hence $J_1(\theta_1) < J_1(\tilde{\theta}_1^*) < J_1(\overset{\circ}{\theta}_1)$. Considering the statistical delay tradeoff characterized in Lemma 1, we must have the associated $J_2(\theta_2) > J_2(\tilde{\theta}_2^*) > J_2(\overset{\circ}{\theta}_2)$, and hence $\theta_2 > \tilde{\theta}_2^* > \overset{\circ}{\theta}_2$. Note that with the characterizations in [15, Lemma 2], we can obtain the following inequality

$$-\frac{1}{\tilde{\theta}^*} \log \mathbb{E}_{z_1} \{e^{-\tilde{\theta}_1^* C_1}\} = -\frac{1}{\tilde{\theta}_1^*} \left(\log \mathbb{E}_{z_2} \{e^{-\theta_2 C_2}\} + \log \mathbb{E}_{z_1} \{e^{(\theta_2 - \tilde{\theta}_1^*) C_1}\} \right) \quad (86)$$

$$< -\frac{1}{\tilde{\theta}_1^*} \left(\log \mathbb{E}_{z_2} \{e^{-\tilde{\theta}_2^* C_2}\} + \log \mathbb{E}_{z_1} \{e^{(\tilde{\theta}_2^* - \tilde{\theta}_1^*) C_1}\} \right) \quad (87)$$

since the RHS of (86) is always greater than the LHS for all $\theta \in [0, \theta_2]$ with given $\tilde{\theta}_1^*$. That is, the condition in (72) is satisfied at $\tilde{\theta}_1^*$. This violates the definition of $\overset{\circ}{\theta}_1$, which is the smallest solution to (72).

Combining the above discussions, we arrive at the conclusion that there is no other θ_1 that can achieve higher effective capacity than (82). Hence, it is indeed the largest achievable constant arrival rate in this case. ■

The aforementioned discussions show the existence of the solution to (30) under the statistical delay constraints. To show the uniqueness, we need the following lemma.

Lemma 3: Consider the function

$$f(\theta_1) = J_2(\theta_2) - J_1(\theta_1) - \log \mathbb{E}_{z_1} \{e^{(\theta_2 - \theta_1) C_1}\}, \text{ for } \theta_1 \leq \underline{\theta}_1 \quad (88)$$

where $(\theta_1, \theta_2) \in \Omega_\varepsilon$. If the following condition

$$\left. \frac{dJ_2(\theta)}{d\theta} \right|_{\theta=\underline{\theta}_2} \leq \left. \frac{dJ_1(\theta)}{d\theta} \right|_{\theta=\underline{\theta}_1} \quad (89)$$

is satisfied, where $(\underline{\theta}_1, \underline{\theta}_2)$ is defined in (70), then $f(\theta_1)$ is increasing in θ_1 .

Proof: Following the proof in Appendix B, we view θ_2 as a function of θ_1 . Now taking the first

derivative of f over θ_1 , we have

$$\frac{df(\theta_1)}{d\theta_1} = \frac{dJ_2(\theta_2)}{dJ_1(\theta_1)} \frac{dJ_1(\theta_1)}{d\theta_1} - \frac{dJ_1(\theta_1)}{d\theta_1} - \frac{\mathbb{E}_{z_1} \{e^{(\theta_2-\theta_1)C_1} C_1\} \left(\frac{d\theta_2}{d\theta_1} - 1\right)}{\mathbb{E}_{z_1} \{e^{(\theta_2-\theta_1)C_1}\}} \quad (90)$$

$$\begin{aligned} &= \frac{dJ_2(\theta_2)}{dJ_1(\theta_1)} \frac{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1} C_1\}}{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\}} - \frac{d\theta_2}{d\theta_1} \frac{\mathbb{E}_{z_1} \{e^{(\theta_2-\theta_1)C_1} C_1\}}{\mathbb{E}_{z_1} \{e^{(\theta_2-\theta_1)C_1}\}} \\ &\quad + \frac{\mathbb{E}_{z_1} \{e^{(\theta_2-\theta_1)C_1} C_1\}}{\mathbb{E}_{z_1} \{e^{(\theta_2-\theta_1)C_1}\}} - \frac{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1} C_1\}}{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\}}. \end{aligned} \quad (91)$$

where $\frac{dJ_1(\theta_1)}{d\theta_1} = \frac{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1} C_1\}}{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\}}$ is substituted into (91).

First, similar to Lemma 2, we can show that the function $g(\theta_2) = \log \mathbb{E}_{z_1} \{e^{(\theta_2-\theta_1)C_1}\}$ is convex in θ_2 , i.e., $\frac{d^2 g(\theta_2)}{d\theta_2^2} \geq 0$. This tells us that the derivative of $g(\theta_2)$ is increasing in θ_2 , and

$$\left. \frac{dg(\theta_2)}{d\theta_2} \right|_{\theta_2=0} = \frac{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1} C_1\}}{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\}}. \quad (92)$$

Therefore,

$$\frac{\mathbb{E}_{z_1} \{e^{(\theta_2-\theta_1)C_1} C_1\}}{\mathbb{E}_{z_1} \{e^{(\theta_2-\theta_1)C_1}\}} - \frac{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1} C_1\}}{\mathbb{E}_{z_1} \{e^{-\theta_1 C_1}\}} \geq 0. \quad (93)$$

Considering the definition of $(\underline{\theta}_1, \underline{\theta}_2)$ in (70), we know that for all $(\theta_1, \theta_2) \in \Omega_\varepsilon$ with $\theta_1 \leq \underline{\theta}_1$, we have $\theta_2 \geq \underline{\theta}_2$. Note that $J_1(\theta_1)$ and $J_2(\theta_2)$ are concave functions according to Lemma 2, i.e., their first derivatives decreases with θ_1 and θ_2 , respectively. Therefore, we have

$$\left. \frac{dJ_1(\theta)}{d\theta} \right|_{\theta=\theta_1} \geq \left. \frac{dJ_1(\theta)}{d\theta} \right|_{\theta=\underline{\theta}_1}, \quad (94)$$

$$\left. \frac{dJ_2(\theta)}{d\theta} \right|_{\theta=\theta_2} \leq \left. \frac{dJ_2(\theta)}{d\theta} \right|_{\theta=\underline{\theta}_2}, \quad (95)$$

which, after combining with the assumption in (89), gives us

$$\frac{dJ_1(\theta_1)}{d\theta_1} \geq \frac{dJ_2(\theta_2)}{d\theta_2}. \quad (96)$$

Next, recalling the statistical delay tradeoff characterized in Lemma 1, we can see that $d\theta_2 < 0$ for $d\theta_1 > 0$, i.e., θ_2 decreases as we increase θ_1 . Then, we can get from (96) that

$$\frac{d\theta_2}{d\theta_1} \leq \frac{dJ_2(\theta_2)}{dJ_1(\theta_1)}. \quad (97)$$

Note that both $\frac{d\theta_2}{d\theta_1}$ and $\frac{dJ_2(\theta_2)}{dJ_1(\theta_1)}$ are negative values. Considering the expression in (91), we now have

$$\frac{df(\theta_1)}{d\theta_1} \geq \left(1 - \frac{dJ_2(\theta_2)}{dJ_1(\theta_1)}\right) \left(\frac{\mathbb{E}_{z_1}\{e^{(\theta_2-\theta_1)C_1}C_1\}}{\mathbb{E}_{z_1}\{e^{(\theta_2-\theta_1)C_1}\}} - \frac{\mathbb{E}_{z_1}\{e^{-\theta_1 C_1}C_1\}}{\mathbb{E}_{z_1}\{e^{-\theta_1 C_1}\}}\right) \geq 0. \quad (98)$$

That is, $f(\theta_1)$ is an increasing function in θ_1 . ■

Note that after eliminating the denominator of both sides of the equation (30), and moving the LHS of the obtained equation to the right side, we can obtain the function given in (88), which is increasing in θ_1 for $\theta_1 \leq \underline{\theta}_1$. Therefore, the solution to the equation (30) is unique.

Case III: Assume $\theta_{1,th} < \theta_{2,th}$. For this case, at $\theta_{1,th}$, we know that

$$R_1 = \frac{J_1(\theta_{1,th})}{\theta_{1,th}} = \frac{J_{th}(\varepsilon)}{\theta_{1,th}} > \frac{J_{th}(\varepsilon)}{\theta_{2,th}} = \frac{J_2(\theta_{2,th})}{\theta_{2,th}} = R_2. \quad (99)$$

The queue at the relay becomes the bottleneck. We need to be careful about the effective capacity in this case. To improve the system performance, we may instead increase the queueing constraint θ_1 at the source, and correspondingly, the queueing constraint θ_2 at the relay can be less. Actually, relaxing the queueing constraint at the source node will not improve the performance, as will be justified later.

First, according to Lemma 2, we can see that as $J_1(\theta_1)$ increases from $J_{th}(\varepsilon)$ to ∞ , θ_1 increases from $\theta_{1,th}$ to ∞ . To the opposite behavior, θ_2 decreases from $\theta_{2,th}$ to some finite value $\theta_{2,0}$, which is the solution to $J_2(\theta) = J_0$. Therefore, from the continuity of θ_2 as a function of θ_1 , we again have one point $(\underline{\theta}_1, \underline{\theta}_2) \in \Omega_\varepsilon$ such that

$$\underline{\theta}_1 = \underline{\theta}_2 \quad (100)$$

and for all $\theta_1 < \underline{\theta}_1$, we have $\theta_1 < \underline{\theta}_1 = \underline{\theta}_2 < \theta_2$. Also, we know that R_1 decreases from $\frac{J_{th}(\varepsilon)}{\theta_{1,th}}$ to $TB \log_2(1 + \text{SNR}_1 z_{1,\min})$, while R_2 increases from $\frac{J_{th}(\varepsilon)}{\theta_{1,th}}$ to some finite value $\frac{J_0}{\theta_{2,0}}$. Therefore, there must be a pair $(\theta_1, \theta_2) \in \Omega_\varepsilon$ such that

$$R_1 = \frac{J_1(\theta_1)}{\theta_1} = R = \frac{J_2(\theta_2)}{\theta_2} = R_2 \quad (101)$$

with the associated statistical queueing constraints denoted as $\check{\theta}_1$ and $\check{\theta}_2$, respectively. For all $\theta_1 < \check{\theta}_1$, we have

$$R_1 = \frac{J_1(\theta_1)}{\theta_1} > \frac{J_2(\theta_2)}{\theta_2} = R_2. \quad (102)$$

Note that the above result implicitly assume that $TB \log_2(1 + \text{SNR}_1 z_{1,\min}) < \frac{J_0}{\theta_{2,0}}$. If this condition does not hold, then θ_1 can take any value, and the only delay is introduced by the queue at the relay node. Hence, the effective capacity under the statistical delay constraint is given by

$$R_\varepsilon(\varepsilon, D_{\max}) = \frac{J_0}{\theta_{1,0}}. \quad (103)$$

Considering the queue stability condition (21), this is possible when the average rate of **R** – **D** link is larger but has more severe fading conditions.

Now, as a stark difference from the previous case, we should have

$$\check{\theta}_1 \geq \underline{\theta}_1. \quad (104)$$

Suppose that $\check{\theta}_1 < \underline{\theta}_1$, we can show the following contradiction. First, at $\check{\theta}_1$, from the definition of $\underline{\theta}_1$ in (100), we have

$$\check{\theta}_1 < \underline{\theta}_1 = \underline{\theta}_2 < \check{\theta}_2. \quad (105)$$

According to the definition of $\check{\theta}_1$ in (101), we can obtain

$$\frac{J_1(\check{\theta}_1)}{\check{\theta}_1} = \frac{J_2(\check{\theta}_2)}{\check{\theta}_2} \Rightarrow J_1(\check{\theta}_1) < J_2(\check{\theta}_2). \quad (106)$$

On the other hand, according to Lemma 1, we should have

$$J_1(\check{\theta}_1) > J_1(\theta_{1,th}) = J_{th}(\varepsilon) = J_2(\theta_{2,th}) > J_2(\check{\theta}_2) \quad (107)$$

leading to contradiction.

Since $\check{\theta}_1 > \underline{\theta}_1$, with (100), we can see that

$$\check{\theta}_1 > \underline{\theta}_1 = \underline{\theta}_2 > \check{\theta}_2. \quad (108)$$

Now, the effective capacity $R_E(\check{\theta}_1, \check{\theta}_2)$ specializes into **Case I** of Theorem 2, we have

$$R_E(\check{\theta}_1, \check{\theta}_2) = \min \left\{ \frac{J_1(\check{\theta}_1)}{\check{\theta}_1}, \frac{J_2(\check{\theta}_2)}{\check{\theta}_2} \right\} = \frac{J_1(\check{\theta}_1)}{\check{\theta}_1} = \frac{J_2(\check{\theta}_2)}{\check{\theta}_2}. \quad (109)$$

Next, we can show the following result.

Proposition 3: The effective capacity in this case is given by

$$R_\varepsilon(\varepsilon, D_{\max}) = \sup_{(\theta_1, \theta_2) \in \Omega} R_E(\theta_1, \theta_2) = R_E(\check{\theta}_1, \check{\theta}_2) = \frac{J_2(\check{\theta}_2)}{\check{\theta}_2} = \frac{J_1(\check{\theta}_1)}{\check{\theta}_1}. \quad (110)$$

Proof: From Proposition 1, we know that

$$R \leq \min \left\{ \frac{J_1(\theta_1)}{\theta_1}, \frac{J_2(\theta_2)}{\theta_2} \right\}. \quad (111)$$

Now, for $\theta_1 > \check{\theta}_1$, we can see that

$$R_1 = \frac{J_1(\theta_1)}{\theta_1} < \frac{J_1(\check{\theta}_1)}{\check{\theta}_1} = R_\varepsilon(\varepsilon, D_{\max}) \quad (112)$$

and for $\theta_1 < \check{\theta}_1$, we have $\theta_2 > \check{\theta}_2$, and hence

$$R_2 = \frac{J_2(\theta_2)}{\theta_2} < \frac{J_2(\check{\theta}_2)}{\check{\theta}_2} = R_\varepsilon(\varepsilon, D_{\max}). \quad (113)$$

Therefore, $R_\varepsilon(\varepsilon, D_{\max})$ in (110) is the largest achievable constant rate in this case. ■

REFERENCES

- [1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013 - 2018.
- [2] A. Goldsmith, *Wireless Communications*, 1st ed. Cambridge University Press, 2005.
- [3] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queuing networks," *IEEE Trans. Auto. Control*, vol. 39, no. 5, pp. 913-931, May 1994.
- [4] F. Kelly, "Notes on effective bandwidth," in *Stochastic Networks: Theory and Applications* Royal Statistical Society Lecture Notes Series, 4. Oxford University Press, 141-168, 1996.
- [5] C.-S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," In *Proceedings of IEEE Infocom*, pp. 1001-1009, 1995.
- [6] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol.2,no. 4, pp.630-643. July 2003
- [7] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp.2318-2328, June 2008.
- [8] A. A. Khalek, C. Caramanis, and R.W. Heath, "Delay-constrained video transmission: Quality-driven resource allocation and scheduling," *IEEE J. Sel. Topics in Sig. Process.*, vol. 9, no. 1, pp. 60-75, Jan. 2015.
- [9] A. Balasubramanian and S.L. Miller, "The effective capacity of a time division downlink scheduling system," *IEEE Trans. Commun.*, vol. 58, no. 1, pp. 73-78, Jan. 2010.
- [10] D. Qiao, M. C. Gursoy, and S. Velipasalar, "Transmission strategies in multiaccess fading channels with statistical QoS constraints," *IEEE Trans. Inform. Theory*, vol. 58, no. 3, pp. 1578- 1593, Mar. 2012.
- [11] Y. Wang and K. J. R. Liu, "Statistical delay QoS protection for primary users in cooperative cognitive radio networks," *IEEE Commun. Letters*, vol. 19, no. 5, pp. 835-838, May 2015.

- [12] J. Tang and X. Zhang, "Cross-layer resource allocation over wireless relay networks for quality of service provisioning," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 645-656, May 2007.
- [13] L. Liu, P. Parag, and J.-F. Chamberland, "Quality of service analysis for wireless user-cooperation networks," *IEEE Trans. Inform. Theory*, vol. 53, no. 10, pp. 3833-3842, Oct. 2007.
- [14] P. Parag, and J.-F. Chamberland, "Queueing analysis of a butterfly network for comparing network coding to classical routing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1890-1907, Apr. 2010.
- [15] D. Qiao, M. C. Gursoy, and S. Velipasalar, "Effective capacity of two-hop wireless communication systems," *IEEE Trans. Inform. Theory*, vol. 59, no. 2, pp. 873- 885, Feb. 2013.
- [16] K.T. Pan and L.-N. Tho, "Effective capacity of dual-hop networks with a concurrent buffer-aided relaying protocol", in Proc. IEEE International Conf. Commun. (ICC 2014), Sydney, NSW, June 2014.
- [17] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks", *Mobile Networks and Applications*, vol. 11, no. 1, pp. 527 - 536, 2005.
- [18] A. A. Khalek and Z. Dawy, "Energy-efficient cooperative video distribution with statistical QoS provisions over wireless networks", *IEEE Trans. Mobile Comp.*, vol. 11, no. 7, pp. 1223 -1236, July 2012.
- [19] Q. Du and C. Zhang, "Queueing analyses and statistically bounded delay control for two-hop green wireless relay transmissions," *Concurrency And Computation: Practice And Experience*, vol. 25, no. 9, pp. 1050-1063, Jun. 2013, DOI: 10.1002/cpe.2875.
- [20] X. Lei, X. Chen, R. Q. Hu, and G. Wu, "Mobile association for heterogeneous wireless relay networks with statistical QoS guarantees," in Proc. IEEE Global Commun. Conf. (Globecom 2013), Atlanta, GA, Dec. 2013.
- [21] B. Xia, Y. Fan, J. Thompson, and H. V. Poor, "Buffering in a three-node relay networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4492-4496, Nov. 2008.
- [22] N. Zlatanov and R. Schober, "Buffer-aided relaying with adaptive link selection-fixed and mixed rate transmission," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 2816-2840, May 2013.
- [23] V. Jamali, N. Zlatanov, and R. Schober, "Bidirectional buffer-aided relay networks with fixed rate transmission-part ii: delay-constrained case," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1339-1355, Mar. 2015.
- [24] T. Charalambous, N. Nomikos, I. Krikidis, D. Vouyioukas, and M. Johansson, "Modeling buffer-aided relay selection in networks with direct transmission capability," *IEEE Commun. Letters*, vol. 19, no. 4, pp. 649-653, Apr. 2015.
- [25] M. Kashef and A. Ephremides, "Optimal partial relaying for energy-harvesting wireless networks," to appear in *IEEE/ACM Trans. Network*.
- [26] A. Ephremides and B. Hajek, "Information theory and communication networks: an unconsummated union," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2416-2434, June 1998.
- [27] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Trans. Inform. Theory*, vol. 50, no. 12, pp. 3062- 3080, Dec. 2004.
- [28] M. Jain, J. I. Choi, T. M. Kim, D. Bharadia, S. Seth, K. Srinivasan, P. Levis, and S. Katti, and P. Sinha, "Practical, real-time, full duplexing wireless," in Proc. ACM Mobile Computing and Networking (MobiCom), Sep. 2011.
- [29] D. Bharadia, E. McMillin, and S. Katti, "Full duplex radio," in Proc. ACM SIGCOMM conf. Appl. Technol., Archit., Protocols Comput. Commun., Hong Kong, Aug. 2013.
- [30] C.-S. Chang, *Performance Guarantees in Communication Networks*, New York: Springer, 1995.